TECHNICAL REPORT ON TEACHER BEHAVIOR, STUDENT ACHIEVEMENT, AND HEAD TEACHER PERFORMANCE

ESRA-PAKISTAN,
PROFESSIONAL DEVELOPMENT

NOVEMBER 2005

TABLE OF CONTENTS

ACKNOWLEDGEMENTS 2

BACKGROUND 3

METHODOLOGY 3

RESULTS 6

SUMMARY 25

ACKNOWLEDGEMENTS

The Evaluation Unit of the Professional Development component of ESRA would like to thank the Government of the Islamic Republic of Pakistan for its assistance on this project. Specifically, the National Education Assessment System (NEAS) office in Islamabad and the Provincial Education Assessment Centres (PEACE) in Balochistan, Sindh, and Lahore have been highly supportive. ESRA would like to continue collaborating with these assessment offices and centres in the future.

The evaluation unit would also like to express its appreciation to all of the field observers who spent time in hundreds of schools collecting data under sometimes challenging conditions, such as remote locations, floods, and extreme temperatures.

Finally, the evaluation unit would like to thank USAID for its support of this initiative. If not for USAID's emphasis on assessment activities and in particular the importance it places on measuring changes in teacher behavior and student achievement, these efforts would not have taken place.

BACKGROUND

The professional development component of the Education Sector Reform Assistance project in Pakistan is responsible for coordinating in-service training for teachers and head teachers in two provinces, Sindh and Balochistan. In the design and implementation of the project, it is essential to address whether the teachers and their students, along with the head teachers, have substantially benefited from the training.

Assessment specialists in the ESRA (AIR) offices in Pakistan and Washington, DC are responsible for setting up the evaluation designs, collecting the data, and analyzing the results of pretest (baseline) and posttest measures. This report discusses the process by which quantitative information on indicators has been collected and analyzed. It also presents the results from the analysis. It is a follow-up document to the earlier report submitted to USAID on teacher, student, and head teacher indicators (*Teacher, Student, and Head Teacher Performance*, ESRA-Pakistan Professional Development, October 2005).

METHODOLOGY

Objectives

The first objective of the quantitative evaluation model is to measure whether teacher behavior has changed as a result of in-service training. The second objective is to examine whether the students taught by those teachers have shown increases in their academic achievement. The third objective is to look at whether head teacher performance has changed as a result of in-service training. The methodology and results sections of this report are organized according to these objectives.

Location

Characteristics of the two provinces (Sindh and Balochistan) where the training is taking place are as follows:

Sindh
- Approximately 39,500 schools, 142,000 teachers, and 2,258,251 students
- School year is from April-March (changed to August-May as of 2005-2006)

Balochistan
- Approximately 9813 schools, 30,000 teachers, and 441,000 students
- School year is from January to December (with some exceptions)

The project intends to provide in-service training for 34,000 elementary school teachers including 3,700 head teachers in the two provinces during the course of the project.

Teacher Behavior

Teacher behavior is measured twice each school year, at the beginning and end of the project-supported in-service training, for each new cohort of intervention teachers, as well as for an

appropriate control group of teachers. The evaluation uses a pre-test/post-test, intervention/control group design. Samples of intervention teachers are selected from each cohort. Samples of control teachers are selected from districts adjacent to the intervention districts in the target provinces.

Teacher behavior is measured using a Teacher Quality Index (TQI), originally provided by USAID. The TQI has 8 items, with one of the items divided into 2 sections, for a total of 9 items. Each item is measured using a four-point Likert scale, giving a total of 36 points. The TQI is shown in Appendix A.

In the initial cohorts (or cycles), samples of 300 schools per province (600 total schools), with one 4th grade (or multi-grade) teacher per school, were observed using the TQI instrument in October-November 2004. There were two observers per teacher. A total of 580 of the teachers were designated as intervention and 20 teachers were control.

An analysis of the design conducted in December 2004-January 2005 found problems in two main areas. First, the school years for the two provinces were different. In Sindh, the school year started in April, so a May pretest would have been recommended. However, the evaluation had not started at that time, so the November pretest was used, even though it was several months into the school year. With the school year was scheduled to end in March, a posttest administration was held in February. In Balochistan, the school year started in January, so the November pretest was too late for any subsequent posttest, thus not permitting a follow-up analysis of results during the school year or training cycle. A pretest was given in February 2005 to the new cohort. Second, the control group was designed as a non-analytical point of reference, with only 10 schools per province. The control group was deemed too small for analysis.

The evaluation model was redesigned in two main ways. The redesign has been employed beginning with cycle 1 in Balochistan (2005 school year) and cycle 2 in Sindh (2005-2006 school year).

- The pretest-posttest schedule was changed to October-May in Sindh and February-October in Balochistan.[1]
- The sample size was changed so that the intervention group had 300 schools and the control group had 150 schools (providing an analytical point of reference).

Student Achievement

Similarly to teacher behavior, student achievement is measured twice each school year, at the beginning and end of the project-supported in-service training, for students taught by each new cohort of intervention teachers, as well as for the students taught by the control group of teachers. The evaluation uses a pre-test/post-test, intervention/control group design.

Student achievement is measured through a multiple choice test for 4th graders with 25 items (one point each) in Mathematics and Urdu. Designed by UNESCO, the instruments were

[1] Note that the school year in Sindh was changed in March -April 2005 so that it would no longer be on an April to May cycle but rather on a September to June cycle. Also note that a small number of schools in Balochistan are not on the February to November school year due to climate issues, so these schools are not included in the sampling frame.

adapted for use by the project and administered to students in the classrooms of the sample teachers.

Student assessment measures were administered to a maximum of 20 randomly selected 4th grade students in the 600 schools in October-November 2004. A total of about 3,600 students took the tests in each province (approximately 7,200 total students).

An analysis of the UNESCO instruments was conducted based on the pretest data. Many of the items had low statistical discrimination (i.e., point biserial correlations). However, since the same instruments needed to be used for the pretest and posttest, the original instruments were maintained in both Sindh and Balochistan for cycle 1. In May 2005, during the head teacher assessment, new items were piloted by the ESRA assessment team. These items had been developed in March and April during two ESRA-sponsored workshops for the MOE/PEACE assessment offices in Sindh and Balochistan provinces. Based on the pilot results, modifications to the instruments using a combined set of new items and UNESCO items were finalized in August 2005. The new instruments will be used starting in cycle 2 in both provinces.

Head Teacher Performance

In May 2005, a head teacher assessment was conducted. It was not originally a part of the USAID indicators but was recommended in early 2005. ESRA developed a 16-item checklist (on a 5-point Likert scale) to measure the essential competencies of a head teacher. The competencies were derived from the training manuals of the implementing partners. The design of the study was as follows:

- A sample of 200 head teachers in 2 groups
- One group of head teachers with training (100) and one group without training (100)
- 70 head teachers in each group from Sindh and 30 from Balochistan
- Comparison of project-trained vs. untrained head teachers

Evaluation Designs

Tables 1 and 2 contain the final designs for the teacher observations, student testing, and head teacher observations for cycles 1 and 2. The designs are different in terms of dates for each of the two provinces due to the differences in the school years. The instruments, sample sizes, and activities, however, should be the same (or at least very similar) for each province.

Table 1. Sindh Timetable

Type of Assessment	Cycle 1		Cycle 2	
	Pretest	Posttest	Pretest	Posttest
Teacher Observations	Nov 2004	Mar 2005	Oct 2005	May 2006
Student Assessments	Nov 2004	Mar 2005	Oct 2005	May 2006
Head Teacher Observations		May 2005		Mar 2006

Table 2. Balochistan Timetable

Type of Assessment	Cycle 1		Cycle 2	
	Pretest	Posttest	Pretest	Posttest
Teacher Observations	Feb 2005	Oct 2005	Feb 2006	Oct 2006
Student Assessments	Feb 2005	Oct 2005	Feb 2006	Oct 2006
Head Teacher Observations		May 2005		Mar 2006

Data Analysis

At this point, a full cycle (pretest and posttest) of teacher and student data from cycle 1 in Sindh (2004-2005 school year) are available, as well as head teacher observation data (May 2005) from the two provinces. Thus, the analysis below contains the results from the following:

- Teacher behavior for cycle 1 in Sindh
- Student achievement for cycle 1 in Sindh
- Head teacher performance for cycle 1 in Balochistan and Sindh

RESULTS

Teacher Performance (in Sindh)

For each analysis, the statistical methodology used was an analysis of variance. Pretest and posttest data were compared. Mean scores and statistical significance levels were calculated. Scores were also analyzed based on pass-fail proficiency cut points.

Table 3 presents overall teacher performance on the Teacher Quality Index before and after the in-service training took place. The statistical analyses show that overall teachers in Sindh performed significantly better after the training ($p < .05$). The mean score increased by over 5 points.[2] The reliabilities for the pretest and the posttest were fairly high (0.80 to 0.87).

Table 3. Teacher Behavior Score Summary

	N	# of Items	Total Points	Raw Score Mean	Standard. Deviation	Alpha (Reliability)
Pretest	264	9	36	15.55	3.72	.80
Posttest	264	9	36	21.16	4.24	.87

Table 4 shows the pretest and posttest performance disaggregated by gender. The statistical analyses indicate that female teachers scored significantly higher than male teachers ($p < .05$) However, there was no interaction effect shown in the results ($p < .05$), which means that scores of female and male teachers progressed in the same way. This can be seen in Graph 1

[2] Note that the n-count is 264 instead of 300. There are two reasons for this. One, there were 10 schools designated as control. These schools were deselected. Two, there were 26 teachers that could not be matched up from the pretest to the posttest. These teachers were deselected.

where the two lines represent scores for females and males, respectively, from pretest to posttest.

Table 4. TQI Scores by Gender

	Male			Female		
	N	Raw Score Mean	Std. Dev.	N	Raw Score Mean	Std. Dev.
Pretest	138	14.97	3.52	126	16.20	3.83
Posttest	138	20.63	4.15	126	21.73	4.27

Graph 1. TQI Scores by Gender

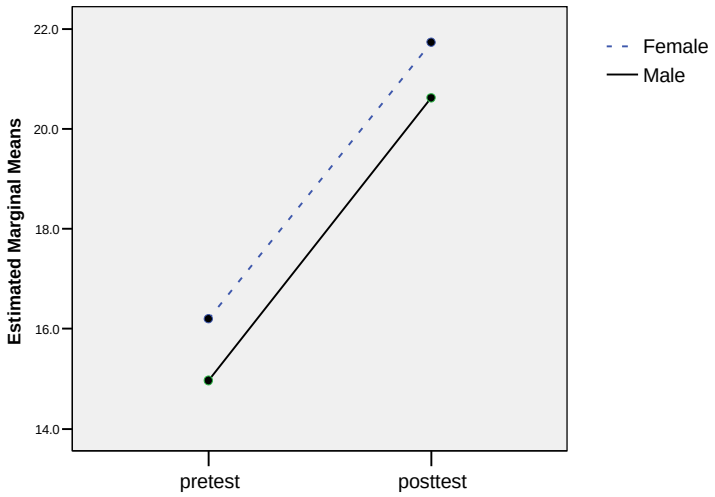


Table 5 presents the teacher performance on each item in both the pretest and the posttest. Score improvements on all items were all statistically significant.

Table 5. TQI Scores by Item

Item		Pretest		Posttest		Stat. Sig. Pretest to Posttest ($p < .05$)
		Mean	Std. Dev.	Mean	Std. Dev.	
1	Prepares the lesson plan in advance.	1.20	.49	2.00	.87	*
2	Implements the lesson plan by giving appropriate attention to student responses and staying focused on the main objective of the lesson.	1.48	.66	2.27	.77	*
3A	Uses different teaching methods during the lesson. (Evaluate each method; mark "N" if method was not used by the teacher.) • Lecture • Teacher-led demonstration • Reading from a book/blackboard	2.01	.65	2.69	.50	*
3B	• Class discussion • Group work • Hands on activities • Student presentation • Role play	1.49	.60	2.24	.72	*
4	Involves students in class activities and encourages interaction among students.	1.68	.75	2.17	.72	*
5	Uses teaching aids (including the blackboard).	1.71	.70	2.29	.72	*
6	Uses the time allocated for the class in an effective manner. (The lesson has a beginning, middle, and an end; it is not repetitive or rushed.)	1.84	.59	2.45	.53	*
7	Has a good command over the subject matter.	2.28	.75	2.77	.54	*
8	Monitors and assesses the students during the lesson. (One example of assessment is to ask open-ended questions to the students; another example, if the students are doing group work, is to go to each group and task appropriate questions.	1.66	.74	2.28	.66	*

Based on an analysis of the score distributions and other data, a total score of 19 was set as a cut point for determining the standard against which teachers would be judged as proficient (or not). Table 6 shows the pre-posttest total score frequencies. About 1 in 5 teachers performed at or above standard before the training (pretest--2004) and about 2 in 3 after the training (posttest--2005). Graphs 2 and 3 are the corresponding histograms. Clearly, the distributions shifted to the right (towards higher scores) from the pretest to the posttest.

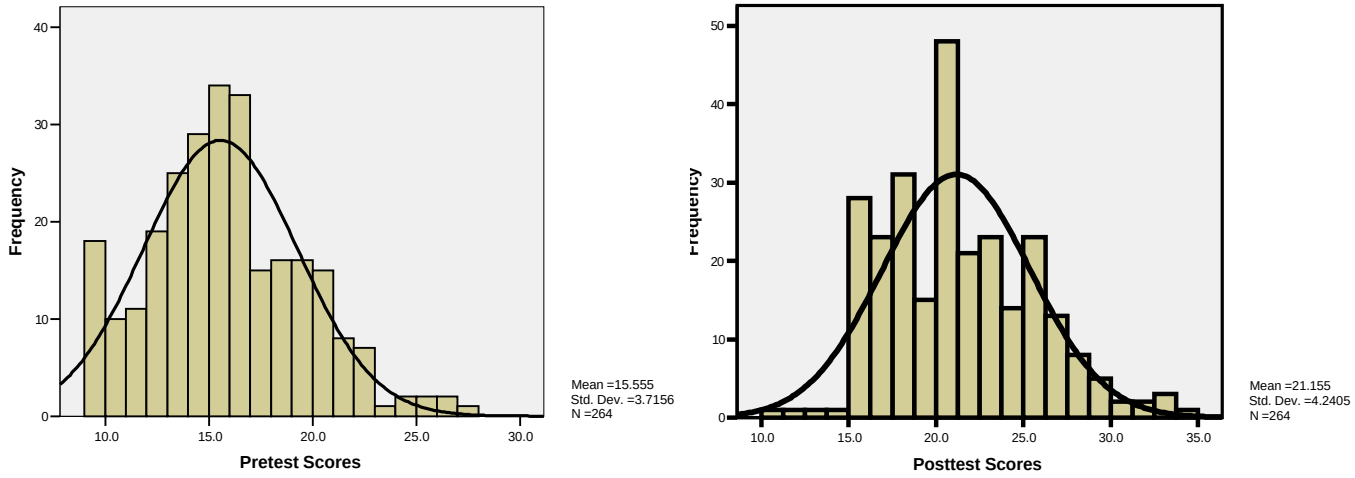
Table 6. TQI Score Frequencies

Total Score	Pretest			Posttest		
	Frequency	Percent	Cum. Percent	Frequency	Percent	Cum. Percent
9.0	13	4.9	4.9	0	0	0
9.5	5	1.9	6.8	0	0	0
10.0	5	1.9	8.7	0	0	0
10.5	5	1.9	10.6	0	0	0
11.0	5	1.9	12.5	1	0.4	0.4
11.5	6	2.3	14.8	0	0	0.4
12.0	7	2.7	17.4	1	0.4	0.7
12.5	12	4.5	22.0	0	0	0.7
13.0	14	5.3	27.3	0	0	0.7
13.5	11	4.2	31.4	1	0.4	1.1
14.0	13	4.9	36.4	0	0	1.1
14.5	16	6.1	42.4	1	0.4	1.5
15.0	15	5.7	48.1	11	4.2	5.7
15.5	19	7.2	55.3	4	1.5	7.2
16.0	20	7.6	62.9	13	4.9	12.1
16.5	13	4.9	67.8	8	3.0	15.2
17.0	8	3.0	70.8	15	5.7	20.8
17.5	7	2.7	73.5	5	1.9	22.7
18.0	11	4.2	77.7	18	6.8	29.5
18.5	5	1.9	79.5	8	3.0	32.6
19.0[3]	12	4.5	84.1	5	1.9	34.5
19.5	4	1.5	85.6	10	3.8	38.3
20.0	12	4.5	90.2	28	10.6	48.9
20.5	3	1.1	91.3	11	4.2	53.0
21.0	6	2.3	93.6	9	3.4	56.4
21.5	2	0.8	94.3	5	1.9	58.3
22.0	5	1.9	96.2	16	6.1	64.4
22.5	2	0.8	97.0	7	2.7	67.0
23.0	0	0	97.0	8	3.0	70.1
23.5	1	0.4	97.3	8	3.0	73.1
24.0	1	0.4	97.7	9	3.4	76.5

[3] Proficiency Cut Score

Total score	Pretest			Posttest		
	Frequency	Percent	Cumulative Percent	Frequency	Percent	Cumulative Percent
24.5	1	0.4	98.1	5	1.9	78.4
25.0	2	0.8	98.9	11	4.2	82.6
25.5	0	0	98.9	2	0.8	83.3
26.0	1	0.4	99.2	10	3.8	87.1
26.5	1	0.4	99.6	1	0.4	87.5
27.0	0	0	99.6	12	4.5	92.0
27.5	0	0	99.6	4	1.5	93.6
28.0	1	0.4	100.0	2	0.8	94.3
28.5	0	0	100.0	2	0.8	95.1
29.0	0	0	100.0	4	1.5	96.6
29.5	0	0	100.0	1	0.4	97.0
30.0	0	0	100.0	2	0.8	97.7
31.5	0	0	100.0	2	0.8	98.5
32.5	0	0	100.0	1	0.4	98.9
33.0	0	0	100.0	2	0.8	99.6
34.0	0	0	100.0	1	0.4	100.0
Total	264	100.0		264	100.0	

Graphs 2 & 3. TQI Frequency Distributions



Student Achievement (in Sindh)

Table 7 presents the summary of results for student performance on the Mathematics and Urdu assessments in Sindh. The statistical analysis results show that students performed significantly better in the posttest than in the pretest for each subject ($p < .01$).[4] The test reliabilities (coefficient alpha) were relatively low; generally, alpha values of 0.80 and above could be expected from such tests.[5]

Table 7. Student Scores by Subject Summary

	N	# of Items	Mean	Std. Deviation	Alpha (Reliability)
Mathematics					
Pretest	2,718	25	6.94	3.18	.54
Posttest	2,718	25	8.98	3.59	.60
Urdu					
Pretest	2,671	25	8.43	3.84	.67
Posttest	2,671	25	10.69	4.59	.76

Table 8 presents the pre- and posttest statistics separated by gender. For Mathematics, the statistical analysis results indicate that female students scored significantly higher than male students on the posttest ($p < .01$). There was an interaction effect shown in the results ($p < .01$), which meant that scores of female and male students changed in different ways. In other words, the males and females had similar scores on the pretest but the score *increase* for female students was significantly higher than for male students. This can be seen in Graph 4, where the two lines represent scores for females and males, respectively, from pretest to posttest.

Table 8. Student Scores by Gender

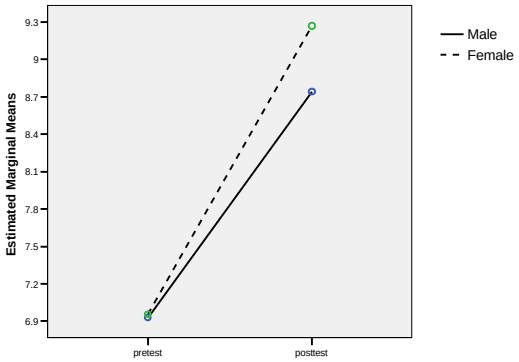
	Male			Female		
	N	Raw Score Mean	Std. Dev.	N	Raw Score Mean	Std. Dev.
Mathematics						
Pretest	1497	6.93	3.10	1221	6.96	3.28
Posttest	1497	8.74	3.39	1221	9.27	3.79
Urdu						
Pretest	1479	7.79	3.57	1192	9.22	4.01
Posttest	1479	9.95	4.34	1192	11.61	4.72

[4] As with the teacher n-counts, the student n-counts are below the original figures. The reason for the decrease is that approximately 20% of the students from the pretest were not in class when the posttest was administered. Discussions with local staff showed that such attrition rates are normal in Pakistan.

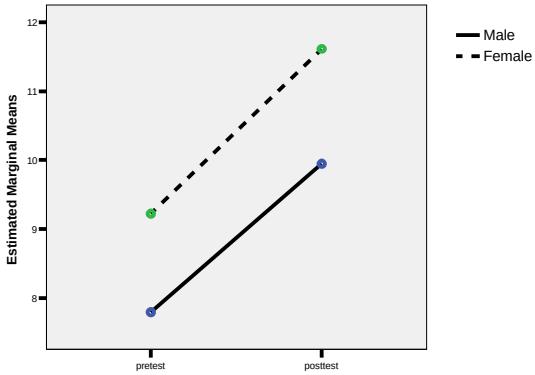
[5] As mentioned in the methodologies section, the UNESCO instruments have been revised. The new items should perform better during cycle 2; also, the number of items on each test has been raised to 30 —in general, having more items on a test leads to greater reliability, all other factors being equal.

Graphs 4 & 5. Student Scores by Gender

Math



Urdu



Also in Table 8, for the Urdu test, female students scored significantly higher than male students on the posttest ($p < .01$). There was no interaction effect shown in the results ($p < .01$), which meant that scores of female and male students changed over time, but in the same way. The differences between the mean scores of female and male students on the pretest and posttest were similar. This is shown by the two lines in Graph 5, which show increasing scores but are nearly parallel.

Tables 9 and 10 present item statistics for Mathematics and Urdu, respectively. For almost all the items for both subjects, the percentage of students answering the item correctly (p-value) increased in the posttest administration. The point biserial coefficients, an indicator of discrimination, are low for many of the items on both the pretest and posttest in both subjects (i.e., below a generally accepted threshold value of 0.25).

Table 9. Mathematics Item P-values and Point Biserials

Item #	Pretest		Posttest	
	P-value[6]	Pt-Biserial[7]	P-value	Pt-Biserial
1	0.44	0.11	.52	0.16
2	0.29	0.11	.29	0.19
3	0.19	0.27	.25	0.23
4	0.21	0.03	.24	0.06
5	0.24	0.12	.29	0.06
6	0.33	0.31	.41	0.35
7	0.38	0.32	.48	0.36
8	0.31	0.27	.44	0.31
9	0.30	0.31	.44	0.34
10	0.24	-0.01	.31	0.06
11	0.34	0.10	.38	0.09
12	0.34	0.19	.38	0.12
13	0.17	0.06	.21	0.08
14	0.29	0.08	.40	0.17
15	0.26	0.09	.30	0.11
16	0.23	0.14	.31	0.23
17	0.32	0.16	.36	0.22
18	0.15	0.10	.18	0.18
19	0.20	0.08	.27	0.14
20	0.32	0.16	.43	0.23
21	0.26	0.11	.39	-0.01
22	0.46	0.24	.60	0.23
23	0.21	0.13	.30	0.21
24	0.29	0.07	.42	0.14
25	0.22	0.28	.40	0.23

[6] The proportion of learners who answered the item correctly.

[7] The correlation between the score on the item and the score on the total instrument; it is a measure of how well the item differentiates between those that answer the i tem correctly or incorrectly and have a high or low total test score respectively.

Table 10. Urdu Item P-values and Point Biserials

Item #	Pretest		Posttest	
	P-value	Pt-Biserial	P-value	Pt-Biserial
1	0.50	0.08	0.59	0.20
2	0.41	0.23	0.46	0.26
3	0.33	0.31	0.34	0.35
4	0.37	0.13	0.42	0.20
5	0.56	0.32	0.68	0.42
6	0.47	0.35	0.57	0.38
7	0.34	0.25	0.44	0.34
8	0.47	0.24	0.54	0.29
9	0.15	0.02	0.21	0.16
10	0.43	0.13	0.49	0.22
11	0.28	0.09	0.39	0.13
12	0.21	0.06	0.31	0.18
13	0.36	0.31	0.47	0.40
14	0.31	0.31	0.35	0.38
15	0.30	0.20	0.44	0.29
16	0.30	0.31	0.44	0.35
17	0.45	0.36	0.56	0.41
18	0.36	0.28	0.49	0.32
19	0.25	0.13	0.30	0.18
20	0.19	0.09	0.30	0.15
21	0.32	0.21	0.34	0.17
22	0.47	0.30	0.58	0.32
23	0.14	0.31	0.28	0.45
24	0.35	0.21	0.42	0.26
25	0.14	0.34	0.25	0.41

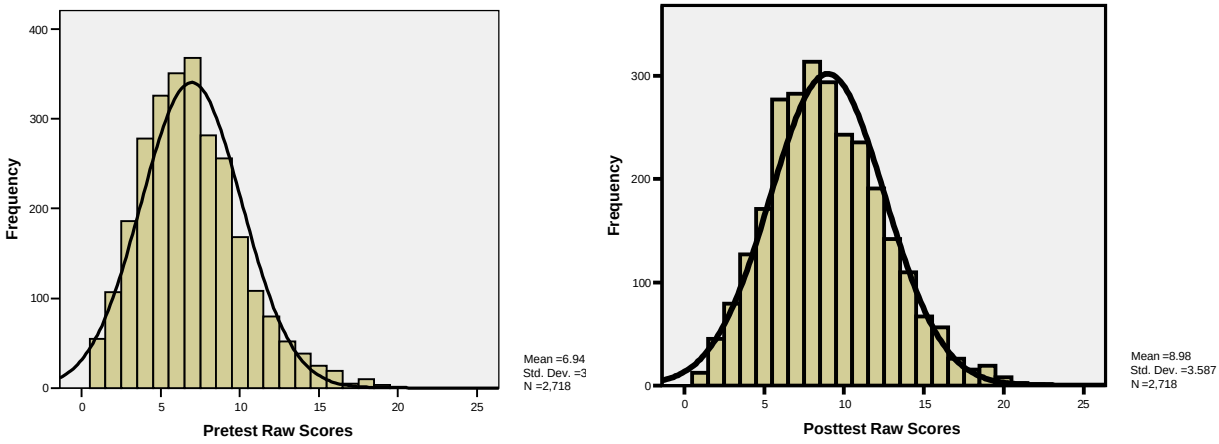
Based on an analysis of the score distributions and other data, a score of 13 for Mathematics assessment was set as a cut point for determining the standard against which students would be judged as proficient (or not). In addition to using the raw scores for the analysis, the student data were scaled for additional analyses , both at the present time and in the future. Using item response theory (IRT), the raw scores were converted using a non-linear transformation into ability scores (or thetas). The ability scores were then converted using a linear transformation into scale scores. The scales were set at 100 (minimum) to 500 (maximum), with passing cut scores of 300. Table 11 shows the Mathematics scores, i.e., the pre- and -posttest total score frequencies. Graphs 6 and 7 are the corresponding histograms. As with the teacher distributions, there was a shift to the right (towards higher scores) from pretest to posttest, although the shift was not as pronounced for the students as it was for the teachers. For Mathematics, the percentage of students obtaining proficiency score increased from 5.6 to 16.4 from the pretest to the posttest.

A score of 15 for Urdu assessment was set as a cut point for determining the proficiency standard. The Urdu data were also scaled based on the same method used for Mathematics data. Table 12 shows the Urdu scores, i.e., the pre- and -posttest total score frequencies separated by gender. Graphs 8 and 9 are the corresponding histograms, again showing a shift to the right. For Urdu, the percentage proficient increased from 7.4 to 21.3.

Table 11. Mathematics Frequencies

Raw Score	Theta	Scale Score	Pretest			Posttest		
			Freq.	Percent	Cum. Percent	Freq.	Percent	Cum. Percent
0	-4.52	100	0	0	0	0	0	0
1	-3.28	100	55	2.0	2.0	12	0.4	0.4
2	-2.54	100	107	3.9	6.0	45	1.7	2.1
3	-2.08	100	186	6.8	12.8	79	2.9	5.0
4	-1.74	130	278	10.2	23.0	127	4.7	9.7
5	-1.45	157	326	12.0	35.0	171	6.3	16.0
6	-1.21	179	351	12.9	47.9	277	10.2	26.2
7	-0.99	200	368	13.5	61.5	283	10.4	36.6
8	-0.79	219	282	10.4	71.9	314	11.6	48.1
9	-0.61	236	256	9.4	81.3	294	10.8	58.9
10	-0.43	252	168	6.2	87.5	243	8.9	67.9
11	-0.25	269	108	4.0	91.4	236	8.7	76.6
12	-0.08	285	80	2.9	94.4	191	7.0	83.6
13[8]	0.08	300	52	1.9	96.3	142	5.2	88.8
14	0.25	316	38	1.4	97.7	110	4.0	92.9
15	0.43	333	25	0.9	98.6	67	2.5	95.3
16	0.61	350	19	0.7	99.3	56	2.1	97.4
17	0.79	366	5	0.2	99.5	26	1.0	98.3
18	0.99	385	10	0.4	99.9	15	0.6	98.9
19	1.21	406	3	0.1	100.0	19	0.7	99.6
20	1.45	428	1	0.0	100.0	8	0.3	99.9
21	1.73	454	0	0	100.0	2	0.1	100.0
22	2.08	487	0	0	100.0	1	0.0	100.0
23	2.54	500	0	0	100.0	0	0	100.0
24	3.28	500	0	0	100.0	0	0	100.0
25	4.52	500	0	0	100.0	0	0	100.0
Total			2,718	100.0		2,718	100.0	

Graphs 6 & 7. Mathematics Frequency Distributions

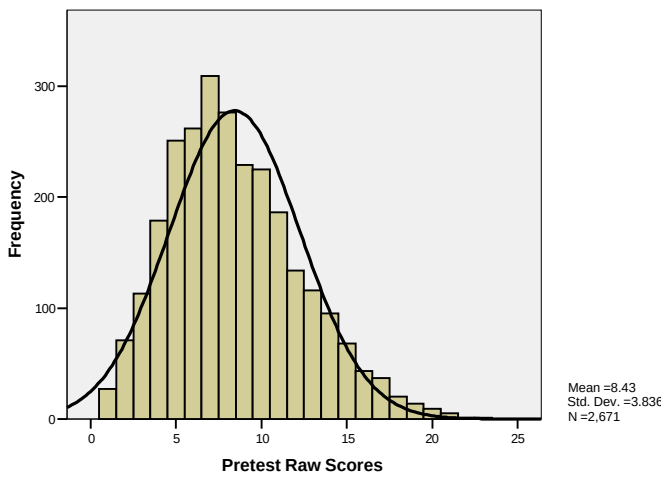


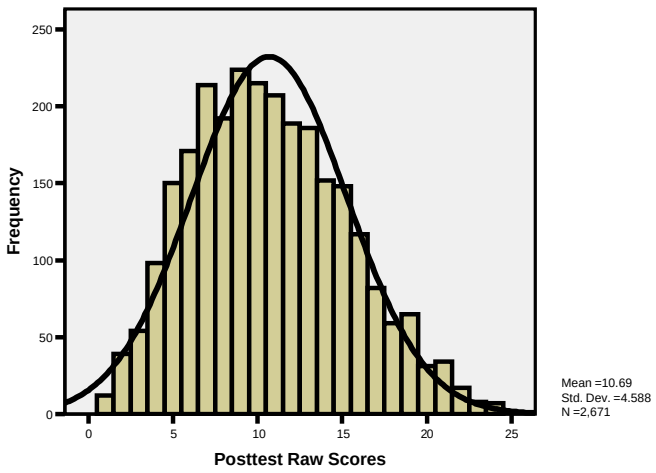
[8] Proficiency Cut Score

Table 12. Urdu Frequencies

Total Score	Theta	Scale Score	Pretest			Posttest		
			Freq.	Percent	Cum. Percent	Freq.	Percent	Cum. Percent
0	-4.58	100	0	0	0	0	0	0
1	-3.33	100	27	1.0	1.0	12	0.4	0.4
2	-2.58	100	71	2.7	3.7	39	1.5	1.9
3	-2.12	100	113	4.2	7.9	54	2.0	3.9
4	-1.77	100	179	6.7	14.6	98	3.7	7.6
5	-1.49	118	251	9.4	24.0	150	5.6	13.2
6	-1.24	142	262	9.8	33.8	171	6.4	19.6
7	-1.02	162	309	11.6	45.4	214	8.0	27.6
8	-0.81	182	276	10.3	55.7	192	7.2	34.8
9	-0.62	200	229	8.6	64.3	224	8.4	43.2
10	-0.44	217	225	8.4	72.7	215	8.0	51.3
11	-0.26	234	186	7.0	79.7	207	7.7	59.0
12	-0.09	250	134	5.0	84.7	189	7.1	66.1
13	0.09	267	116	4.3	89.0	186	7.0	73.0
14	0.26	283	95	3.6	92.6	152	5.7	78.7
15[9]	0.44	300	68	2.5	95.1	148	5.5	84.3
16	0.62	317	43	1.6	96.7	117	4.4	88.7
17	0.81	335	37	1.4	98.1	82	3.1	91.7
18	1.02	355	20	0.7	98.9	59	2.2	93.9
19	1.24	375	14	0.5	99.4	65	2.4	96.4
20	1.49	399	9	0.3	99.7	31	1.2	97.5
21	1.77	425	5	0.2	99.9	34	1.3	98.8
22	2.12	458	1	0.0	100.0	17	0.6	99.4
23	2.58	500	1	0.0	100.0	8	0.3	99.7
24	3.33	500	0	0	100.0	7	0.3	100.0
25	4.57	500	0	0	100.0	0	0	100.0
Total			2,671	100.0		2,671	100.0	

Graphs 8 & 9. Urdu Frequency Distributions





[9] Proficiency Cut Score

Head Teacher Performance (in Balochistan and Sindh)

As mentioned in the methods section, the head teacher assessment was different from the teacher behavior and student achievement assessments in three main ways. One, there was no pretest, only a posttest. Two, there was a viable control group. Three, the assessment was conducted in both Balochistan and Sindh provinces. Hence, the comparisons are for the intervention vs. control group and for the two provinces. Scores are the result of the administration of a head teacher instrument with 16 items and a 5-point scale (80 points maximum).

Table 13 provides a summary of the results for the trained and untrained teachers by province and by gender. There was a statistically significant difference between the scores of the trained and untrained head teachers. The difference in the mean scores was slightly over 3 points. By province, there was a significant difference in favor of the trained head teachers in Balochistan but not in Sindh. By gender, there was a significant difference in the mean scores for male head teachers (in favor of the trained head teachers) but not for the female head teachers. The reliability of the head teacher instrument was very high, at 0.94.

Table 13. Head Teacher Summary

	Trained Head Teachers			Untrained Head Teachers			Stat. Sig. ($p < .05$)
	N	Mean	Std. Dev.	N	Mean	Std. Dev.	
Overall	98	46.60	8.54	101	43.21	10.26	*
By Province							
Balochistan	28	45.68	9.22	32	39.00	1.43	*
Sindh	70	46.97	8.29	69	45.16	10.62	
By Gender							
Male	62	46.79	7.78	50	42.48	10.26	*
Female	36	46.28	9.81	51	43.92	10.30	

In Table 14, an analysis of the results by item show several significant differences between the trained and untrained head teachers. Overall, trained teachers significant higher than untrained teachers on the following individual measures:

Management style
Awareness regarding rules and regulations
Awareness regarding curriculum
Mastery over the subject matter
Guidance and counseling
Dealing with the parents
Ability of teamwork
Ability of classroom observation
Mentoring ability

Table 14. Head teachers by Item

Item	Trained Head Teachers		Untrained Head Teachers		Stat. Sig. ($p < .05$)
	Mean	Std. Dev.	Mean	Std. Dev.	
Management style	3.39	.78	3.09	.81	*
Awareness regarding rules & regulations	3.08	.73	2.83	.81	*
Maintenance of record	3.28	.78	3.07	.89	
Academic supervision	2.88	.71	2.83	.95	
Awareness regarding curriculum	2.51	.78	2.11	.90	*
Mastery over the subject matter	2.88	.68	2.65	.85	*
IT skills	1.39	.64	1.32	.75	
Guidance and counseling	2.91	.71	2.63	.98	*
Dealing with the parents	2.96	.85	2.67	1.02	*
Ability of planning	2.73	.88	2.62	.96	
Ability of team work	3.10	.82	2.74	.95	*
Ability of classroom observation	2.96	.69	2.75	.78	*
Attitude with the teachers	3.45	.72	3.30	.76	
Attitude with the students	3.37	.71	3.31	.72	
Monitoring ability	2.93	.79	2.73	.90	
Mentoring ability	2.76	.77	2.49	.93	*

In Table 15, for Balochistan, trained teachers perform significantly higher than untrained teachers on the following individual measures:

Awareness regarding rules & regulations
Maintenance of record
Awareness regarding curriculum
Mastery over the subject matter
Ability of team work
Ability of classroom observation
Monitoring ability
Mentoring ability

Table 15. Balochistan Head Teachers by Item

Item	Trained Head Teachers		Untrained Head Teachers		Stat. Sig. ($p < .05$)
	Mean	Std. Dev.	Mean	Std. Dev.	
Management style	3.25	.75	2.88	.94	
Awareness regarding rules & regulations	3.11	.69	2.56	.80	*
Maintenance of record	3.25	.65	2.69	.82	*
Academic supervision	2.75	.70	2.38	.79	
Awareness regarding curriculum	2.18	.91	1.38	.66	*
Mastery over the subject matter	2.86	.65	2.34	.79	*
IT skills	1.54	.69	1.25	.72	
Guidance and counseling	2.71	.66	2.41	.88	
Dealing with the parents	2.68	.77	2.34	1.00	
Ability of planning	2.79	.83	2.41	.88	
Ability of team work	3.18	.86	2.59	.88	*
Ability of classroom observation	2.93	.66	2.41	.56	*
Attitude with the teachers	3.43	.63	3.28	.73	
Attitude with the students	3.39	.63	3.28	.68	
Monitoring ability	2.89	.74	2.50	.76	*
Mentoring ability	2.71	.66	2.19	.74	*

In Table 16, for Sindh, trained teachers perform significantly higher than untrained teachers on only one measure, Management style.

Table 16. Sindh Head Teachers by Item

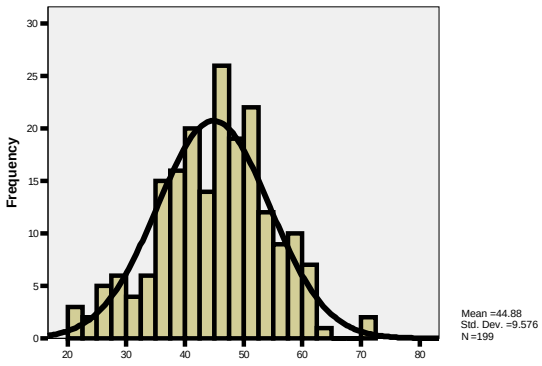
Item	Trained Head Teachers		Untrained Head Teachers		Stat. Sig. ($p < .05$)
	Mean	Std. Dev.	Mean	Std. Dev.	
Management style	3.44	.79	3.19	.73	*
Awareness regarding rules & regulations	3.07	.75	2.96	.79	
Maintenance of record	3.29	.84	3.25	.86	
Academic supervision	2.93	.71	3.04	.95	
Awareness regarding curriculum	2.64	.68	2.45	.80	
Mastery over the subject matter	2.89	.69	2.80	.85	
IT skills	1.33	.61	1.35	.76	
Guidance and counseling	2.99	.71	2.74	1.01	
Dealing with the parents	3.07	.86	2.83	1.00	
Ability of planning	2.71	.90	2.72	.98	
Ability of team work	3.07	.80	2.81	.97	
Ability of classroom observation	3.97	.70	2.91	.82	
Attitude with the teachers	3.46	.76	3.30	.77	
Attitude with the students	3.36	.74	3.32	.74	
Monitoring ability	2.94	.81	2.84	.95	
Mentoring ability	2.77	.82	2.62	.99	

Tables 17 and 18, along with Graphs 10, 11, and 12, provide information on the frequencies of the head teacher performance scores. Notice that the frequency distribution for the trained head teachers is centered at approximately 50 while the frequency distribution for the untrained head teachers is centered at closer to 40.

Table 17. Head Teacher Frequencies

Score	Frequency	Percent	Cumulative Percent
20	2	1.0	1.0
22	1	.5	1.5
23	1	.5	2.0
24	1	.5	2.5
25	1	.5	3.0
27	4	2.0	5.0
28	4	2.0	7.0
29	2	1.0	8.0
30	3	1.5	9.5
32	1	.5	10.1
33	3	1.5	11.6
34	3	1.5	13.1
35	2	1.0	14.1
36	6	3.0	17.1
37	7	3.5	20.6
38	7	3.5	24.1
39	9	4.5	28.6
40	3	1.5	30.2
41	10	5.0	35.2
42	7	3.5	38.7
43	6	3.0	41.7
44	8	4.0	45.7
45	6	3.0	48.7
46	13	6.5	55.3
47	7	3.5	58.8
48	9	4.5	63.3
49[10]	10	5.0	68.3
50	8	4.0	72.4
51	9	4.5	76.9
52	5	2.5	79.4
53	4	2.0	81.4
54	8	4.0	85.4
55	5	2.5	87.9
56	1	.5	88.4
57	3	1.5	89.9
58	6	3.0	93.0
59	4	2.0	95.0
60	3	1.5	96.5
61	3	1.5	98.0
62	1	.5	98.5
63	1	.5	99.0
72	2	1.0	100.0
Total	199	100.0	

Graph 10: Head Teacher Frequency Distribution

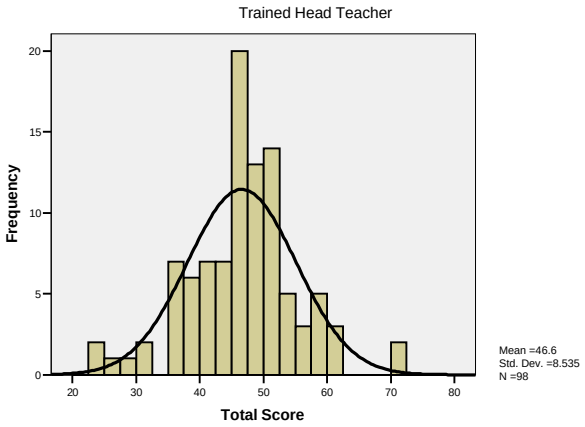


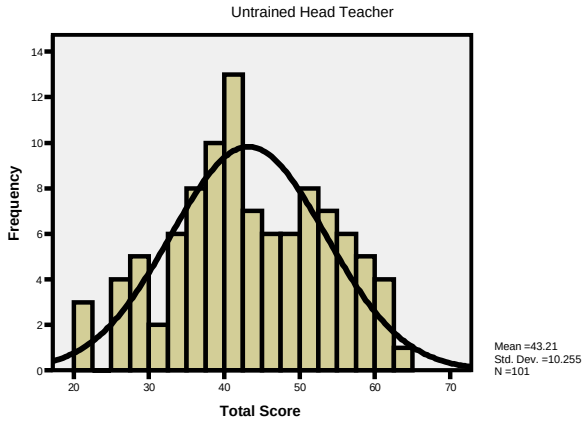
[10] Cut point for "Satisfactory"

Table 18. Head Teacher Frequencies by Trained/Untrained

Total Score	Trained Head Teacher			Untrained Head Teacher		
	Frequency	Percent	Cum. Percent	Frequency	Percent	Cum. Percent
20	0	0	0	2	2	2
22	0	0	0	1	1	3
23	1	1	1	0	0	3
24	1	1	2	0	0	3
25	1	1	3.1	0	0	3
27	0	0	3.1	4	4	6.9
28	1	1	4.1	3	3	9.9
29	0	0	4.1	2	2	11.9
30	1	1	5.1	2	2	13.9
32	1	1	6.1	0	0	13.9
33	0	0	6.1	3	3	16.8
34	0	0	6.1	3	3	19.8
35	0	0	6.1	2	2	21.8
36	3	3.1	9.2	3	3	24.8
37	4	4.1	13.3	3	3	27.7
38	0	0	13.3	7	6.9	34.7
39	6	6.1	19.4	3	3	37.6
40	1	1	20.4	2	2	39.6
41	1	1	21.4	9	8.9	48.5
42	5	5.1	26.5	2	2	50.5
43	2	2	28.6	4	4	54.5
44	5	5.1	33.7	3	3	57.4
45	3	3.1	36.7	3	3	60.4
46	12	12.2	49	1	1	61.4
47	5	5.1	54.1	2	2	63.4
48	5	5.1	59.2	4	4	67.3
49	8	8.2	67.3	2	2	69.3
50	6	6.1	73.5	2	2	71.3
51	5	5.1	78.6	4	4	75.2
52	3	3.1	81.6	2	2	77.2
53	1	1	82.7	3	3	80.2
54	4	4.1	86.7	4	4	84.2
55	3	3.1	89.8	2	2	86.1
56	0	0	89.8	1	1	87.1
57	0	0	89.8	3	3	90.1
58	3	3.1	92.9	3	3	93.1
59	2	2	94.9	2	2	95
60	0	0	94.9	3	3	98
61	3	3.1	98	0	0	98
62	0	0	98	1	1	99
63	0	0	98	1	1	100
72	2	2	100	0	0	100
Total	98	100		101	100	

Graphs 11 & 12. Head Teacher Frequency Distributions by Trained/Untrained





SUMMARY

The following points summarize the findings of the cycle 1 assessments:

Teacher Behavior

- Overall, teachers performed significantly better after the training than before the training.
- Female teachers performed significantly better than male teachers both before and after the training.
- The TQI instrument had good reliability.

Student Achievement

Mathematics

- Overall, students performed significantly better on the posttest than on the pretest.
- Female students scored significantly higher than male students on the posttest.
- The score increase for female students was significantly higher than male students.

Urdu

- Overall, students performed significantly better on the posttest than on the pretest.
- Female students scored significantly higher than male students on the pretest and posttest.
- The score increases for female and male student were not significantly different.
- Both instruments had low test reliabilities and low point-biserials on many of the items, suggesting the need to revise the instruments (which has already taken place).

Head Teacher Performance

- Overall, trained head teachers performed significantly better than untrained head teachers.
- For Balochistan, trained head teachers performed significantly better than untrained teachers. For Sindh, there was no significant difference.
- Male trained head teachers performed significantly better than untrained male head teachers. There was no difference between female trained and untrained head teachers.
- Instrument reliability was very high.